Evolution of Tokenization: Handling Out-of-Vocabulary Words

Input Text:

"The antidisestablishmentarianism was extraordinary"

1. Traditional Word-Level
(vocab_size = 5,000)

The	[UNK]	was	[UNK]

50% UNK rate
Information loss

2. Character-Level
(no vocabulary limit)

0% UNK rate
Very long sequences

3. Subword Tokenization
(WordPiece/BPE)

| The | anti | ##dis | ##esta | ##bl | was | The | rare_word | z |
| ##sh | ##ment | ##arian | ##ism | | | | | |

0% UNK rate

4. Hybrid Approach
(Subword + Character fallback)

Adaptive handling

Comparison Metrics

Method	UNK Rate	Seq Length	Efficiency	Coverage
Traditional	High	Short	Fast	Poor
Character	None	Very Long	Slow	Perfect
Subword	Minimal	Medium	Good	Excellent
Hybrid	None	Variable	Adaptive	Perfect

Evolution toward better OOV handling

Key Insight
Subword tokenization
balances all factors

Modern Trend
Hybrid approaches for
ultimate flexibility